Research Statement

Xin Wen, Tongji University

I am determined to equip computers with the same visual capabilities as humans (perceiving & reasoning). To build intelligent machines that can not only perceive the visual world, but also tackle challenging reasoning problems under uncertainty, I pursue the answers via studies of **Computer Vision**, **Machine Learning**, and **Interdisciplinary Data Science** (CV & NLP). More specifically, I am interested in designing scalable learning algorithms to form structured representations of the visual world in multiple modalities to analyze massive video datasets with complex structures. In particular, I eager to advance methods in the following research areas:

- Structured Video Understanding
- Multi-Modal/Self-Supervised Representation Learning
- Long-Term Video Semantic Understanding

Besides, I am also open to other relevant interesting topics, such as Image Processing, 3D Vision, and Computer Graphics, *etc*. More information about my research, projects and biography is available at `http://wen-xin.info`.

1 Background & Research Plan

I would like to first introduce my work in Video Retrieval, conducted during my internship at ByteDance AI Lab, followed by my research plan in this topic.

1.1 My Work on Video Retrieval

I have conducted continuous research in the field of Video Retrieval, which from my perspective is a straightforward task for evaluating the quality of video representations. The central task of Video Retrieval is to predict the similarity between video pairs. Current approaches mainly follow two schemes: to compute similarity using video-level representations (first scheme) or frame-level representations (second scheme). For methods using video-level representations, early studies typically employ code books [1, 2, 3] or hashing functions [4, 5] to form video representations, while later approach (Deep Metric Learning [6]) is introduced to generate video representations by aggregating the pre-extracted frame-level representations. In contrast, the approaches following the second scheme typically extract frame-level representations to compute frame-to-frame similarities, which are then used to obtain video-level similarities [7, 8, 9, 10]. With more elaborate similarity measurements, they typically outperform those methods with the first scheme.

For both schemes, the frames of a video are commonly processed as individual images or short clips, making the identification of informative frames difficult. As the visual scene of videos can be redundant (such as Scenery Shots or B-rolls), potentially unnecessary visual data may dominate the video representation, and mislead the model to retrieve negative samples sharing similar scenes. Motivated by the effectiveness of self-attention mechanism in capturing long-range dependencies [11], we propose to incorporate temporal information between frame-level features using self-attention mechanism, helping the model focus on more informative frames, thus obtaining more relevant and robust features.

To supervise the optimization of video retrieval models, current state-of-the-art methods [6, 9] commonly perform instance discrimination on pair-wise labels with triplet loss [12]. However, the relation that triplets can cover is limited, and the performance of triplet loss is highly subject to the time-consuming hard-negative sampling process [13]. Inspired by the recent success of contrastive learning on self-supervised learning [14, 15] and the nature of video retrieval datasets that rich negative samples are readily available, we propose a supervised contrastive learning method for video retrieval. With the help of a shared memory bank, large quantities of negative samples are utilized efficiently with no need for manual hard-negative sampling. Furthermore, by conducting gradient analysis, the property of automatic hard-negative mining is also discovered in the proposed method.

Extensive experiments are conducted on multi video retrieval datasets, such as FIVR [16], CC_WEB_VIDEO [17] and EVVE [18]. The proposed method shows a significant performance advantage (*e.g.* $\sim$ 17% mAP on FIVR-200K) over state-of-the-art methods with video-level features, and deliver competitive results with a much lower computational cost when compared with methods using frame-level features.

1.2 Potential Future Work

TL;DR To utilize information from additional multi-modalities for video retrieval.

The research focus on Content-Based Video Retrieval has shifted from Near-Duplicate Video Retrieval (NDVR) [17, 19] to Fine-grained Incident Video Retrieval (FIVR) [16], Event-based Video Retrieval (EVR) [18] and Action Video Retrieval (AVR) [20]. Different from NDVR, these tasks are more challenging in terms that they require higher-level representation describing the semantics of relevant incidents, events, and actions.

In above-mentioned work, we tried to tackle this problem through temporal correlation modeling with self-attention mechanism to help the model capture long-range dependencies and concentrate on more informative frames. We demonstrate considerable performance gain in our work, but there is still a long way towards solving video retrieval problem on these datasets. By analyzing a number of the bad cases, we discovered that those videos generally have the problem of low resolution, severe jittering and poor lighting. In such cases, the information obtained from only the visual scenes of the video is extremely limited and hence leads to unsatisfactory performance.

Utilizing information from additional multi-modalities of the videos can be a key to solving this problem. Videos are far more than sequences of frames, they naturally contain rich information in multiple modalities, such as visual scenes, audios, and captions, *etc*. For example, for videos in social media, the corresponding titles, descriptions, tags, and comments may also be available. These associated multi-modal information is complementary to the video itself and can be used to describe the video more comprehensively, helping to learn better representations. In a more task-specific scenario containing many texts such as tags and comments, techniques in tasks (*e.g.* Sentimental Analysis, Word/Paragraph Embedding) of Neural Language Processing (NLP), may be adopted. For the scenario containing additional audios, methods in Automatic Speech Recognition (ASR) can be helpful. According to my experience in handling industrial projects during the internship, the reasonable use of multi-modal features can often give stable performance gains.

1.3 Beyond Video Retrieval

TL;DR To build hidden motion representation with relation networks for human action recognition.

Although CNNs have yielded unprecedented progress on a wide range of image-centric benchmarks, they are also shown to be deficient in modeling long-term semantic dependencies within videos. This problem is not only seen in content-based video retrieval, but also other video-based tasks that require the understanding of pairwise object relation, global context and temporal semantics, such as video object detection [21] and action recognition [22], *etc*. Specifically, it is revealed in [22] that current action recognition methods are typically over-dependent on appearance features, and lack the ability of utilizing temporal information for the modeling of motion features.

A possible direction is to form structured motion representations. Inspired by [23, 24], a video can be viewed as a space-time graph of object proposals. To be specific, each frame of the video is first encoded into a few object proposals, and the relationship between object proposals is modeled with the self-attention mechanism or graph neural networks, then the motion representation of the video is extracted from the relationship between object proposals. Different from the flow-based works that only consider the difference between adjacent frames, this is more similar with the dense trajectories [25] that can better model long-term motion and global context, which has been proved to be important for fine-grained action understanding [22].

2 Survey of Related Work

In recent works of Contend-Based Video Retrieval, frame-level representations are first extracted independently, then aggregated by feature aggregation models to obtain video-level representations (optional) and finally trained with metric learning. Therefore, the related work is introduced from these three aspects: Frame Feature Representation, Feature Aggregation, and Metric Learning.

2.1 Frame Feature Representation

A common strategy is to extract frame-level representations independently as image representations. Early approaches employed handcrafted features including the Scale-Invariant Feature Transform (SIFT) features [26, 27, 17], the Speeded-Up Robust Features (SURF) [28, 7], Colour Histograms in HSV space [29, 30, 5], and Local Binary Patterns (LBP) [31, 32, 33], *etc*.

Deep Convolutional Neural Networks (CNNs) have proved to be versatile representation tools in recent approaches. The application of Maximum Activation of Convolutions (MAC) and its variants [34, 35, 36, 37, 38, 39, 40], which extract frame descriptors from activations of a pre-trained CNN model, have achieved great success in both fine-grained image retrieval and video retrieval tasks [40, 2, 41, 6, 9]. Intermediate Maximum Activation of Convolutions (iMAC) [40] applies MAC to different intermediate layers of a CNN then concatenate them. Regional Maximum Activation of Convolutions (R-MAC) [37] build feature vectors that encode several image regions rather than the whole image, and L_N-iMAC [9] applies R-MAC on the activations of the intermediate convolutional layers, but the regional feature maps are stacked rather than summed. Besides variants of MAC, Sum-Pooled Convolutional features (SPoC) [42] and Generalized Mean (GeM) [43] pooling are also considerable counterparts.

2.2 Feature Aggregation

Typically, the video feature aggregation paradigm can be divided into two categories: (1) local feature aggregation models [44, 45, 46, 47] which are derived from traditional local image feature aggregation models, and (2) sequence models [48, 49, 50, 20, 11, 51] that model the temporal order of the video representation.

The commonly used local feature aggregation models include Bag-of-Words [44, 45], Fisher Vector [46], and Vector of Locally Aggregated Descriptors (VLAD) [47], of which the unsupervised learning of a visual code book is required. The NetVLAD [52] transfers VLAD into a differential version, and the clusters are tuned via back-propagation instead of k-means clustering. NeXtVLAD [53] further decomposes the high-dimensional feature into a group of relatively low-dimensional vectors with attention before applying NetVLAD aggregation over time, which is both effective and parameter efficient. In terms of the sequence models, the Long Short-Term Memory (LSTM) [48] and Gated Recurrent Unit (GRU) [49] are commonly used to model contextual information within a long-range for video re-localization and copy detection [20, 54]. Besides, The effectiveness of self-attention in capturing short and long-range dependency with attention mechanism has been proved with the success of Transformer [11]. For the feature aggregation of videos, this also shows success in video classification [55] and object detection [56], opening new possibilities for feature aggregation for video retrieval.

2.3 Metric Learning

Metric learning aims to learn an embedding that minimizes the distance between related samples and maximizes it between irrelevant ones. Metric learning have been commonly used in face recognition [57, 58, 59], image retrieval [60, 13, 61, 62] and video retrieval [6, 9]. With only pair-wise labels available, the triplet loss [12] is commonly used in video retrieval tasks [6, 9]. The classic approach in [6] performs hard negative mining to generate hard triplets, but despite both the off-line triplet generation stage and the training stage are time-consuming, the information that triplets can convey is limited [13]. Although [63] showed the triplet loss can perform competitively against other popular metric learning approaches with proper hard negative sampling strategy, the proposed PK sampling strategy is only compatible with datasets with class-level labels.

Contrastive learning has become the common training architecture of recent self supervised learning works [64, 65, 66, 14, 15], in which the positive and negative sample pairs are constructed with a pretext task in advance, and the model tries to distinguish the positive sample from massive randomly sampled negative samples in a classification manner. The contrastive loss typically performs better in general than triplet loss on representation tasks [15], as the triplet loss can only handle one positive and negative at a time. The core of the effectiveness of contrastive learning is the use of rich negative samples [66], one approach is to sample them from a shared memory bank [67], and [14] replaced the bank with a queue and used a moving-averaged encoder to build a larger and consistent dictionary on-the-fly. Apart from self-supervised learning, supervised contrastive learning for classification tasks is also discussed in [68], in which a modified batch contrastive loss that supports an arbitrary number of positives is proposed to leverage label information effectively. As we only have pair-wise labels, our supervised contrastive learning approach is more similar to the self-supervised approach, where each anchor is coupled with only one positive.

References

[1] Yang Cai, Linjun Yang, Wei Ping, Fei Wang, Tao Mei, Xian-Sheng Hua, and Shipeng Li. Million-scale near-duplicate video retrieval system. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 837–838, 2011.

[2] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *International conference on multimedia modeling*, pages 251–263. Springer, 2017.

[3] Kaiyang Liao, Hao Lei, Yuanlin Zheng, Guangfeng Lin, Congjun Cao, Mingzhu Zhang, and Jie Ding. Ir feature embedded bof indexing method for near-duplicate video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3743–3753, 2018.

[4] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 423–432, 2011.

[5] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, 15(8):1997–2008, 2013.

[6] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. Near-duplicate video retrieval with deep metric learning. In *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[7] Chien-Li Chou, Hua-Tsung Chen, and Suh-Yin Lee. Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia*, 17(3):382–395, 2015.

[8] Hao Liu, Qingjie Zhao, Hao Wang, Peng Lv, and Yanming Chen. An image-based near-duplicate video retrieval and localization using improved edit distance. *Multimedia Tools and Applications*, 76(22):24435–24456, 2017.

[9] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Visil: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6351–6360, 2019.

[10] Hung-Khoon Tan, Chong-Wah Ngo, Richard Hong, and Tat-Seng Chua. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 145–154, 2009.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[12] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[13] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[16] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Fivr: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia*, 21(10):2638–2652, 2019.

[17] Xiao Wu, Alexander G Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 218–227, 2007.

[18] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Event retrieval in large video collections with circulant temporal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2459–2466, 2013.

[19] Yu-Gang Jiang, Yudong Jiang, and Jiajun Wang. Vcdb: a large-scale database for partial copy detection in videos. In *European conference on computer vision*, pages 357–371. Springer, 2014.

[20] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. Video re-localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018.

[21] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020.

[22] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020.

[23] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[24] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Asian Conference on Computer Vision*, pages 363–378. Springer, 2018.

[25] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4305–4314, 2015.

[26] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501, 2007.

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[28] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[29] Yanbin Hao, Tingting Mu, Richang Hong, Meng Wang, Ning An, and John Y Goulermas. Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, 19(1):1–14, 2016.

[30] Weizhen Jing, Xiushan Nie, Chaoran Cui, Xiaoming Xi, Gongping Yang, and Yilong Yin. Global-view hashing: harnessing global relations in near-duplicate video retrieval. *World wide web*, 22(2):771–789, 2019.

[31] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[32] Lifeng Shang, Linjun Yang, Fei Wang, Kwok-Ping Chan, and Xian-Sheng Hua. Real-time large scale near-duplicate web video retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 531–540, 2010.

[33] Zhipeng Wu and Kiyoharu Aizawa. Self-similarity-based partial near-duplicate video retrieval and alignment. *International Journal of Multimedia Information Retrieval*, 3(1):1–14, 2014.

[34] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.

[35] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good practice in cnn feature transfer. *arXiv preprint arXiv:1604.00133*, 2016.

[36] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.

[37] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[38] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2017.

[39] Omar Seddati, Stéphane Dupont, Saïd Mahmoudi, and Mahnaz Parian. Towards good practices for image retrieval based on cnn features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1246–1255, 2017.

[40] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[41] Yang Li, Yulong Xu, Jiabao Wang, Zhuang Miao, and Yafei Zhang. Ms-rmac: Multiscale regional maximum activation of convolutions for image retrieval. *IEEE Signal Processing Letters*, 24(5):609–613, 2017.

[42] Artem Babenko and Victor Lempitsky. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 1269–1277, 2015.

[43] Yanbin Hao, Tingting Mu, John Y Goulermas, Jianguo Jiang, Richang Hong, and Meng Wang. Unsupervised t-distributed video hashing and its deep hashing extension. *IEEE Transactions on Image Processing*, 26(11):5531–5544, 2017.

[44] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[45] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003.

[46] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[47] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[48] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[49] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[50] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[51] Jin Xia, Jie Shao, Cewu Lu, and Changhu Wang. Weakly supervised em process for temporal localization within video. In *2019 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[52] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[53] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[54] Yaocong Hu and Xiaobo Lu. Learning spatial-temporal features for video copy detection by the combination of cnn and rnn. *Journal of Visual Communication and Image Representation*, 55:21–29, 2018.

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[56] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[58] Qiong Cao, Yiming Ying, and Peng Li. Similarity metric learning for face recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2408–2415, 2013.

[59] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[60] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[61] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, Romain Garnier, and Neil M Robertson. Ranked list loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5207–5216, 2019.

[62] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[63] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[64] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[65] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[66] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[67] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[68] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *ArXiv*, abs/2004.11362, 2020.